Scene text localization based on the ultimate opening

Thomas Retornaz and Beatriz Marcotegui

Centre de Morphologie Mathématique, Ecole des Mines de Paris,
35, rue Saint-Honoré, 77305 Fontainebleau cedex, France
`retornaz,marcotegui@cmm.ensmp.fr`

Abstract Database indexing and retrieval tools can enormously benefit from the automatic detection and processing of textual data in images. We present a new connected-component (CC) based method using morphological numerical residues for automatic localization of text in general image database. This technique tends to be robust to font, scale and slant changes and detects superimposed as well as scene text. The robustness of our approach is proven by the results in *ImagEval* evaluation campaign, which database included old postcards, graphic schemes, stamps, indoor and outdoor scene images and also images without any textual data. In spite of the wide variety of texts and images our approach obtains interesting results without parameter tuning for each image class.

Keywords: scene-text localization, Connected component approach, ultimate opening, ImagEval, indexing.

1. Introduction

Multimedia databases, both personal and professional, are developing at a high rate and the need for automated management tools is now imperative.The effort devoted by the research community to content-based image indexing is also huge, but bridging the semantic gap is difficult: the low level descriptors used for indexing (e.g:interest points, texture descriptors) are not enough for an ergonomic manipulation of big and generic image databases. The text present in a scene is usually linked to the semantic context of the image and constitutes a relevant descriptor for content-based image indexing.

Many algorithms focusing on scene text detection have been designed in the past few years. The reader may refer to [8] and [10] for a complete survey of text detection applications and systems. Basically, text localization approaches can be sorted into two main categories. The first category is texture based (which includes derivative and frequency approaches). Jung [7] use as features the pixel values in a star-like pattern. Clark [3] propose five localized features and combine them to get candidate text regions, and claim that their approach is invariant to scale and orientation.

The frequency domain is also used: Fourier transform [15], discrete cosine transform [4], wavelet [17], multi-resolution edge detector [2]. These methods perform quite well for small characters, because small texts produce strong texture response. Their extension to generic text imposes important constraints on character alignment and/or color distribution together with the introduction of a multiscale strategy. The second category is the connected component approach. Some authors use color quantization process [16], morphological operations [5] or split-and-merge algorithm [11] to obtain candidate CCs. These methods could effectively deal with different types of text but require too many heuristic rules such as aspect ratio of characters, horizontal constraint.

Our work focuses on a very generic image database, as illustrated in Figure 7. It contains a wide variety of texts in terms of size, color, font (even manuscript), complex background and also typical scene text deformations due for example to perspective or non planar support. We will build our system on the three following hypotheses: the text should be readable (contrasted), it should be composed of group of characters, and characters of the same text zone should have similar geometric features.

Our system consists in 5 steps that will be described in corresponding sections hereafter. In Section 2, the non linear scale-space approach based on morphological numerical residues is presented. It allows us to extract candidate CCs using contrast information. Section 3 introduces a two-step filter which removes CCs that are certainly non text, detected on basic and easily computed features. Section 4 gives features for discriminating text CCs from non text ones. Then the Section 5 presents the learning strategy. Finally an alignment and merging analysis to filter out the last remaining non-text CCs is introduced in Section 6. Section 7 describes evaluation issues and overall results in the *ImagEval* evaluation campaign. Section 8 is devoted to our conclusions and perspectives.

2. Ultimate opening

Beucher [1] has recently introduced a non linear scale-space based on morphological numerical residues. It enables us to select CCs by trying to avoid at least a priori on size information. Here is an overview of ultimate openings.

2.1 Reminder on attribute opening and closing

Binary connected opening: We define a 2D binary image as a subset of $\mathbb{Z}^2$. A binary image X can be decomposed into its connected components $C_i \in \mathcal{C}$, where $\mathcal{C}$ is a connectivity class and i some index set I. We can extract a connected component CC to which belongs an element $x \in X$ by using a binary connected opening Γ_x (see [14] for theoretical background).

Criterion and trivial opening: A trivial opening Γ_κ uses an increasing criterion κ to accept or reject a connected set (that κ is increasing means A satisfies κ which implies B satisfies κ for all $B \supseteq A$). Usually κ is in the form

$$\kappa(CC) = (AttributeValue(CC) \geq \lambda) \tag{1}$$

with AttributeValue(CC) some real-valued attribute of CCs (such as area, height,...) and λ the attribute threshold. The trivial opening Γ_κ on a CC simply means that if $\kappa(CC) = True$, we keep CC otherwise we discard it.

Binary Attribute Opening: A binary attribute opening in X consists in a trivial opening of each CC. We can define it as follows:

$$\Gamma^\kappa(X) = \bigcup_{x \in X} \Gamma_\kappa(\Gamma_x(X)) \tag{2}$$

Gray-Scale case: To apply the definitions to gray scale case we can use the basic threshold decomposition process. Fast implementations can be found in [13].

The top-hat associated with attribute opening or closing is a powerful operator to select image structures but a priori knowledge of size of these structures is required. Next we show the interest of some residual operators, in particular ultimate opening/closing.

2.2 Definition of Ultimate Opening/Closing

The ultimate opening θ was introduced by Beucher [1]. This residual operator analyses the evolution of each pixel x of a grayscale image I subject to a family of openings of increasing sizes (with size parameter λ). The difference between two consecutive openings (named residue) is considered and two significant pieces of information are kept for each pixel. $R_\theta(I)$ registers the value of the maximal residue (contrast information) and $q_\theta(I)$ registers the size of the opening that produced the maximal residue (i.e.the size of the structure that contains the considered pixel). Note that this maximum may not be unique. In that case, as proposed by Beucher, we keep the greatest λ value for which this maximum occurs.

The ultimate opening is then defined as follows:

$$\begin{array}{l} R_\theta(I) = sup(r_\lambda(I)), \forall \lambda \geq 1 : \text{ with } r_\lambda(I) = \gamma_\lambda - \gamma_{\lambda+1} \\ q_\theta(x) = max(\lambda) : \lambda \geq 1, r_\lambda(x) = R_\theta(x) \text{ and } > 0 \end{array} \tag{3}$$

Now we denote ν^γ an ultimate opening linked with a family of morphological openings and υ^γ the ultimate closing defined by duality.

Figure 1 illustrates the action of υ^γ on a profile, here we have used closings with a family of linear structuring elements of increasing size.

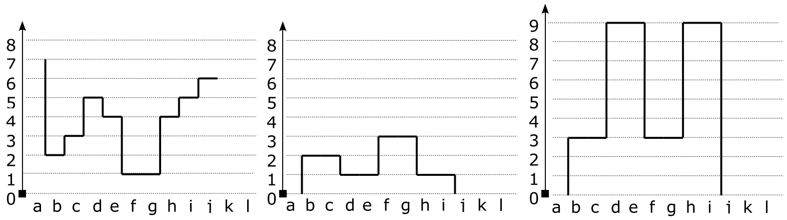


Figure 1. Operator υ^{γ}. From left to right: profile I, R_θ and q_θ

2.3 Ultimate Opening/Closing and attribute

The definitions of ν^{γ}/ υ^{γ} using morphological opening (respectively closing) could be extended by using attribute opening (respectively closing). We denote ν^{κ}/ υ^{κ} these operators and the new definitions of R_θ and q_θ are obtained replacing γ by Γ^{κ} in eq. 3. We can re-use the Figure 1 to illustrate the action of υ^{κ}, we just replace morphological closings by area closings of increasing size.

2.4 Extracting CC using ultimate attribute closing

Different attributes may be considered for the ultimate opening. For text detection, the height of the bounding box containing CCs is the most suited one. The largest opening considered is λ equal to 1/3 of the image height because 1) characters are rarely larger and 2) we avoid artifacts that usually occur at the late steps of the ultimate opening.

The CCs are computed by thresholding the residual image R_θ. First a global low threshold (value fixed at 8) is applied in order to remove really low contrasted structures. Then, a local adaptive threshold is applied on each CC. The aim is to separate CCs possibly merged due to the extremely low threshold previously used. This local threshold is based on the mode of R_θ in each CC. Since ultimate opening tends to give the same R_θ valuation to all the pixels of a contrasted structure, it makes sense to use the R_θ mode. A threshold of mode/2 is applied in each CC. See Figure 2 for an example.

Note that only dark CCs have been extracted. In order to detect also light CCs, the same procedure should be applied to the inverse of the image.

Finally, note also that we only use luma information. ν may be applied to some "color gradient" in order to partially integrate colorimetric information. Due too a significant number of very thin letters in the database (particularly in the old postcard subset), we decide to stop going further.

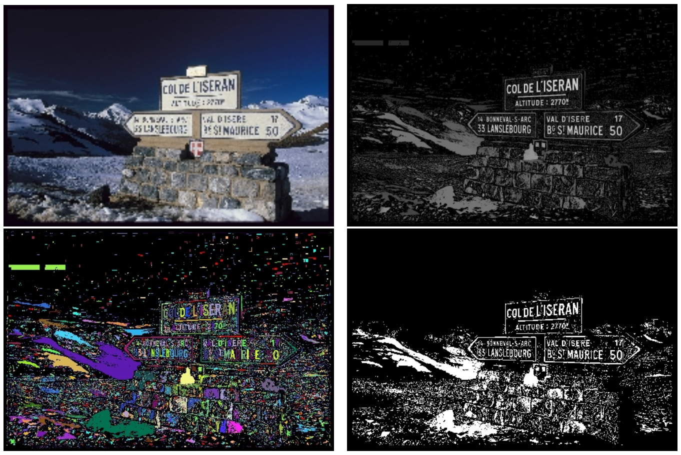


Figure 2. CCs extraction: From left to right and up to down: Original Image I, $R_\theta(I)$ (Gamma Corrected $\gamma = 3$) and $q_\theta(I)$ (randomized), detected CCs.

3. Coarse filtering

In the previous step a lot of CCs were obtained. Most of them do not correspond to characters and can be removed by simple tests as those proposed in this section.

Textured regions typically produce a lot of CCs. They deserve particular attention or a lot of false positives may appear. Therefore we observed the answer of texture to ultimate opening and developed a simple module for texture removal. This module consists in counting the number of CCs grouped by a unitary dilation of the initial CC set. If this number is high (used value 500), the grouped CCs are removed.

A second simple consideration also leads to remove a great number of CCs. Characters are supposed to be made of strokes of constant thickness. We estimate the line thickness of each CC and we remove any CC whose height is smaller than two times its thickness. The thickness is estimated using the distance function in horizontal and vertical directions.

Figure 3 shows that most of non text CCs were successfully discarded at the end of the two-step process.

4. Features to discriminate characters

The number of CCs was considerably reduced in the previous step. We will now characterize the remaining ones in more details. We will classify components in text/non text categories based on the following features.

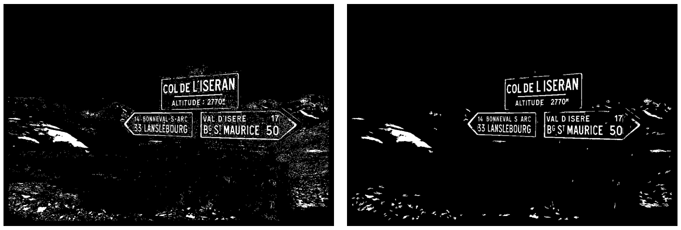


Figure 3. Coarse filtering. Left: result after removing grouped CCs. Right: result after the second step.

Geometric features: The simplest features we can consider for a CC is the height (H_{bb}) and width (W_{bb}) of its bounding box, the area of the CC (A_{cc}), and the area of the bounding box (A_{bb}). The inverse values of (H_{bb}), (W_{bb}) and (A_{cc}) are also introduced, allowing us to compute aspect ratios. Some systems commonly use these features and their combinations to discard non text CCs (especially for superimposed characters). In the context of scene text their characterizing performances fall down. We will combine these features, which assess the scale of the character, with more discriminant features. 7 geometrical features are considered.

Stroke thickness estimation: A character is supposed to be composed of lines of constant thickness. This seems to be a key characteristic for text CC discrimination. However its estimation is not so easy, this is why several estimators are considered. We propose first to compute the max of distance function inside CCs (Max_{cc}), which generally leads to thickness overestimation because of stroke intersections. We also use run-length distribution inside CCs (as Ashida System in [12]), more precisely we compute the distributions (line per line and column per column) of run-length inside CC. Several parameters are selected from these distributions: mean, median and variance of the RLE distribution (RLE_{MedX}, RLE_{MedY}, RLE_{MoyX}, RLE_{MoyY}, RLE_{VarX}, RLE_{VarY}). RLE is generally sensitive to text deformations.

Moreover, we consider the coherence of the stroke thickness. We compute the distributions of differences (between consecutive lines or columns) of run-length inside CCs. We keep the mean and variance of these distributions (ΔRLE_{MoyX}, ΔRLE_{MoyY}, ΔRLE_{VarX}, ΔRLE_{VarY}). 11 features estimating the stroke thickness and consistency are used altogether.

Shape regularity features: Text CCs have more regular shapes than arbitrary CCs. We propose to compute some shape parameters which express this regularity. For example text CCs have a limited number of holes, rather regular contours and important compactness. So to assess these char-

Table 1. Confusion table of our learning system.

Detected	*Characters*	*No Characters*
GT		
Characters	89.1	10.9
No Characters	9.7	90.3

acteristics, we propose the following features: Euler Number E, perimeter P, compacity (A_{cc}/P^2) and complexity (P/A_{cc}) (as [9]). The last three characteristics are also computed after filling the holes of CCs because we think that this transformation could be relevant to discriminate text from non text CCs. (which makes 7 shape regularity features)

Contrast: We estimate the contrast of a CC as the Maximum Inter Class Variance $(M.V.I)$ in its bounding box. We include this feature because the text is supposed to be contrasted in order to be readable. We also include $(M.V.I/\text{Overall Variance})$ which is a normalized feature.

5. Machine learning

We use machine learning techniques in order to achieve the classification of text/non text CCs base on the 27 features presented above. To this purpose we use quadratic Linear Discriminant Analysis (LDA) [6], which attempts to predict the classification as a linear combination of the selected features and cross products.

The classifier was trained on 177000 CCs extracted from 350 images and labeled by hand as text or no text. The classical cross-validation technique has been used to compute the cost function and to evaluate the performances of the designed classifier. The confusion table is shown in Table 1. A misclassification rate of about 10 % is obtained for both text and no text CCs.

Figure 4 presents typical classification results of the designed classifier. CCs of Figure 3 (right) classified as text are shown in Figure 4 (top). Figure 4 (down) shows the result of coarse filtering (left) and the CCs classified as text leading to false positives in the church image (see Figure 7).

6. Grouping characters in text zones

Text zones are rarely composed of a single character. This remark is often used to reduce the number of false positives at the cost of a few false negatives. A commonly accepted constraint is to impose at least 3 characters in order to accept a text zone. We develop a two step alignment and merging analysis.

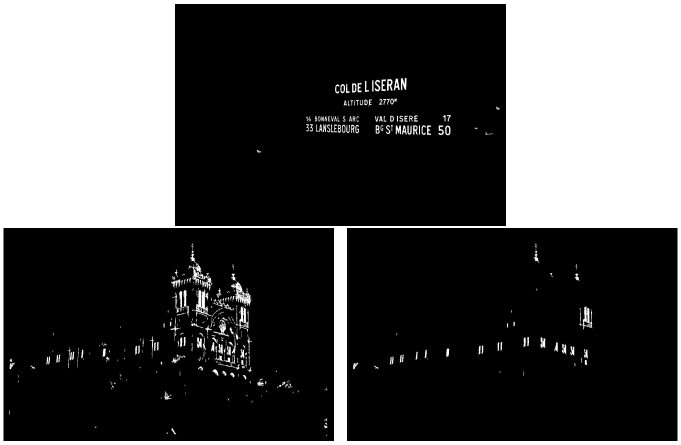


Figure 4. Machine Learning. Top: CCs of Figure 3 (right) classified as text. Bottom left: result of the coarse filtering. Right: CCs classified as text.

Strong layout: We start from the CCs classified as text by the learning stage and we merge CCs if they verify the following constraints (see Figure 5):

- the difference of corresponding bounding boxes heights is smaller than the smallest bounding box height ($|H_1 - H_2| < min(H_1, H_2)$).
- the distance of the bounding boxes centers in the vertical direction is less than 70% of the smallest bounding box height ($\Delta_C y < 0.7min(H_1, H_2)$).
- the distance of the bounding boxes centers in the horizontal direction is less than the smallest bounding box height ($\Delta X < min(H_1, H_2)$).

Only groups of at least 3 characters are preserved.

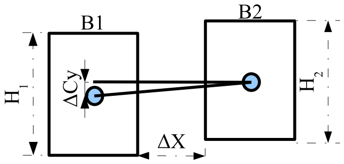


Figure 5. Parameters used in the merging process

Relaxation of merging strategy: The text zones detected in the previous step are considered as seeds and extended in the horizontal direction (see Figure 6 (Left)) by a certain amount (here 200 pixels). A new merging process is applied in these extended areas with the following constraints: 1) the merging criteria are the same that those used for seed creation, 2) groups of at least 2 CCs are accepted and 3) CCs classified as no text by the learning approach take part in the process. If groups of at least 2 CCs are found, they are added to the final detected box (see the '17' string in Figure 6 (Right)). Thus groups of only 2 CCs can be detected, but only if they are in the neighborhood of a seed. Since CCs classified as no text are considered, some misclassification of the learning step are recovered.

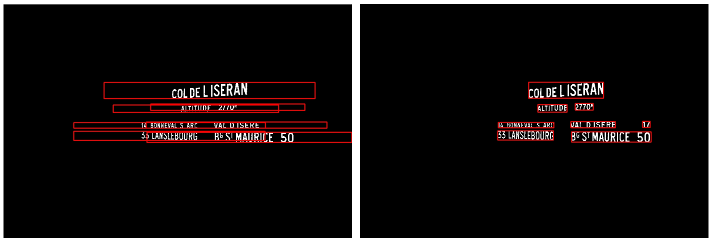


Figure 6. Grouping strategy. Left: resulting seeds after the application of the strong layout with extended boxes for the relaxation step. Right: result after relaxation step.

Final Grouping The *ImagEval* committee established some rules in order to define the ground truth (see www.imageval.org for details). We merge all the boxes verifying the merging criterion defined by the committee.

Finally, given that we detect clear and dark text separately, we merge boxes if they intersect.

7. Experimental Results

7.1 Evaluation issues

The global database contains 500 images including old postcards, graphic schemes, stamps, indoor and outdoor scene images and also images without any textual data. Furthermore images can be grayscale or color. The ground truth database was created by the evaluation committee and was not known by the participants. The committee used the ICDAR[1] metric for performance evaluation (see [12]). This metric is based on the intersection ratio between detected text zones and ground truth. If no text is detected for an image without textual information, both precision and recall are set to 1. The results are given in Section 7.3.

[1] International Conference on Document Analysis and Recognition

7.2 Parameters of the system

Along the paper several parameters have been introduced. Given that the ground truth was unknown, systematic optimization was not possible. All these parameters have been empirically tuned in order to cope with the huge variability of the blind test database. A second database with similar characteristics has been provided for the official test. The robustness of our method is proven by the results we have obtained (see section below).

Note that we provide simply an overall result. In fact it is difficult to quantify the effects of each steps separately because our approach finds CCs, then the ground truth is known at the bounding box level. Even if we have CCs tagged as text, there is still a gap (known as the character restoration step) before we could submit our CCs to an Optical Character Recognition system.

7.3 Overall Results

The results were given to the participants by the *ImagEval* evaluation committee. The database contains 500 images (190 old postcards, 206 color images and 104 black and white images). Table 2 gives the precision, recall and F-mean computed with the ICDAR performance metric [12].

Precision	Recall	F-Mean
0.48	0.65	0.55

Table 2. Overall Performance

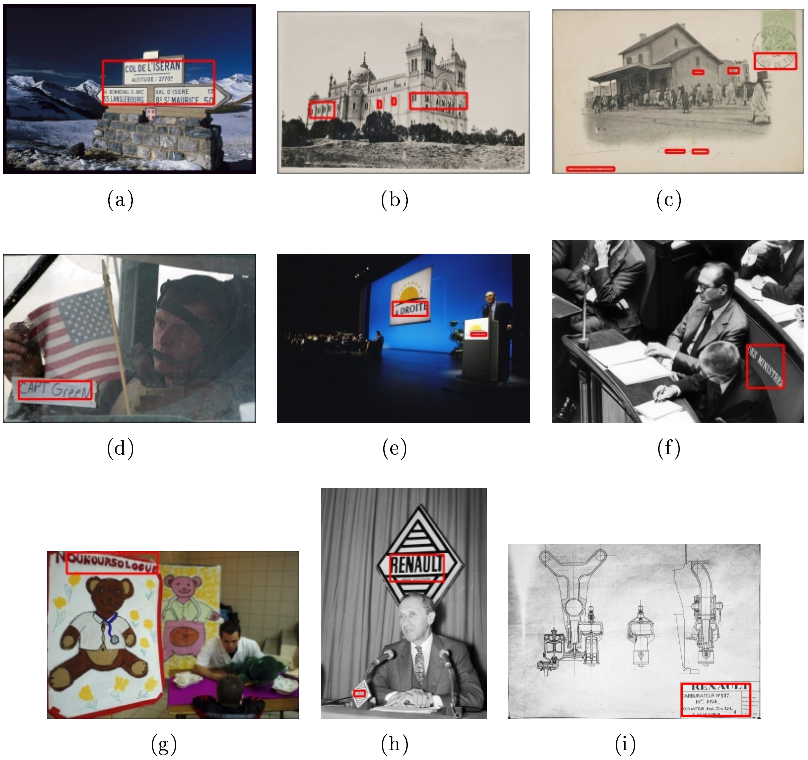
Precision	Recall	F-Mean
0.41	0.57	0.48

Table 3. Overall Performance ICDAR dataset

In Figure 7 the variability of the database and the visual quality of the results can be observed. In general the results are satisfactory and the proposed method seems to cope with the variability of text zones in scale(Figures 7(e),7(h)), font(Figure 7(d)) , and slant(Figure 7(f)). Some false positives are detected, for example in structured zones (such as barriers or balcony handrails). Some false negatives are also present: the merging rules do not take into account vertical text and therefore it is discarded. If the characters are not correctly defined, as it is the case in the postal card, the CC extraction step fails.

We also provide in Table 3, results from *free* ICDAR database [2]. Note that we merge results on TrailTrain and TrialTest subset. All system parameters remain, except for the last merging step (ie **Final Grouping**). We replace the merging criterion defined by *ImagEval* committee by a simple inclusion criterion due to a huge difference on annotation strategy (*ImagEval*

[2] Aviable at http://algoval.essex.ac.uk/icdar/Datasets.html sec: Robust Reading and Text Locating



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 7. Examples of detected text zones

committee annotates at the sentence level whereas ICDAR committee annotates at the word level). No additional efforts were made to improve the results.

8. Conclusions

The whole system has been empirically tuned for the extremely diverse *ImagEval* database without ground truth information. In spite of the wide variety of text and images, a promising score is obtained for the whole database. The ground truth is now available, and we are considering parameter optimization in a probabilistic framework. The strong point of our system is the relative robustness against changes in scale, fonts, complex backgrounds and typical deformations of scene text (perspective, non planar support). The system may be significantly improved for a restricted application by modifying specifically the learning step. In future work we plan to integrate the q_θ (the second piece of information provided by the ultimate opening operator), an important source of information not used yet, in the CC extraction step.

References

[1] Serge Beucher, *Numerical residues*, Mathematical morphology: 40 years on, 2005, pp. 23–32.

[2] X. Chen, J. Yang, J. Zhang, and A. Waibel, *Automatic detection and recognition of signs from natural scenes*, IEEE Transactions on Image Processing **13** (2004January), no. 1, 87–99.

[3] P. Clark and M. Mirmehdi, *Recognising text in real scenes*, International Journal on Document Analysis and Recognition **4** (2002August), no. 4, 243–257.

[4] D. Crandall, S. Antani, and R. Kasturi, *Extraction of special effects caption text events from digital video*, International Journal of Document Analysis and Recognition **5** (2003April), no. 2-3, 138–157.

[5] Y.M.Y. Hasan and L.J. Karam, *Morphological text extraction from images*, IEEE Trans. Image Processing **9** (2000November), no. 11, 1978–1983.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning : Data mining, inference, and prediction*, 1st ed. 2001. Corr. 3rd printing, 2003, Springer Series in Statistics, Springer, New York, 2001.

[7] K. Jung and J. Han, *Hybrid approach to efficient text extraction in complex color images*, Pattern Recognition Letter **25** (2004April), no. 6, 679–699.

[8] K. Jung, K.I. Kim, and A.K. Jain, *Text information extraction in images and video: a survey*, Pattern Recognition **37** (2004May), no. 5, 977–997.

[9] M. León, S. Mallo, and A. Gasull, *A tree structured-based caption text detection approach*, Proc. fifth iasted international conference visualization, imaging and image processing, 2005September, pp. 220–225.

[10] J. Liang, D. Doermann, and H. Li, *Camera-based analysis of text and documents: a survey*, International Journal on Document Analysis and Recognition **7** (2005July), no. 2-3, 84–104.

[11] R. Lienhart and W. Effelsberg, *Automatic text segmentation and text recognition for video indexing*, Multimedia Syst. **8** (2000), no. 1, 69–81.

[12] S.M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, H. Miyao, J. Zhu, W. Ou, C. Wolf, J.M. Jolion, L. Todoran, M. Worring, and X. Lin, *Icdar 2003 robust reading competitions:entries, results, and future directions*, International Journal on Document Analysis and Recognition **7** (2005July), no. 2-3, 105–122.

[13] A. Meijster and M.H.F. Wilkinson, *A comparison of algorithms for connected set openings and closings*, IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002April), 484–494.

[14] Jean Serra, *Image analysis and mathematical morphology: Theorical advances*, Vol. 2, Academic Press, 24-28 OvalRoad, London NW1 7DX, 1988.

[15] B.K. Sin, S.K. Kim, and B.J. Cho, *Locating characters in scene images using frequency features*, International conference on pattern recognition, 2002, pp. III: 489–492.

[16] K. Wang and J.A. Kangas, *Character location in scene images from digital camera*, Pattern Recognition **36** (2003October), no. 10, 2287–2299.

[17] Q. Ye, Q. Huang, W. Gao, and D. Zhao, *Fast and robust text detection in images and video frames*, Image and Vision Computing **23** (2005June), no. 6, 565–576.